

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP2005/017696

International filing date: 27 September 2005 (27.09.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-345392
Filing date: 30 November 2004 (30.11.2004)

Date of receipt at the International Bureau: 17 November 2005 (17.11.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: 2 0 0 4 年 1 1 月 3 0 日

出 願 番 号
Application Number: 特 願 2 0 0 4 - 3 4 5 3 9 2

パリ条約による外国への出願
に用いる優先権の主張の基礎
となる出願の国コードと出願
番号

The country code and number
of your priority application,
to be used for filing abroad
under the Paris Convention, is

J P 2 0 0 4 - 3 4 5 3 9 2

出 願 人
Applicant(s): 松下電器産業株式会社

2 0 0 5 年 1 1 月 2 日

特許庁長官
Commissioner,
Japan Patent Office

中 嶋



【書類名】	特許願
【整理番号】	7048060194
【提出日】	平成16年11月30日
【あて先】	特許庁長官殿
【国際特許分類】	G06F 17/30
【発明者】	
【住所又は居所】	大阪府門真市大字門真 1 0 0 6 番地
【氏名】	松下電器産業株式会社内 稲葉 光昭
【発明者】	
【住所又は居所】	大阪府門真市大字門真 1 0 0 6 番地
【氏名】	松下電器産業株式会社内 菅野 祐司
【特許出願人】	
【識別番号】	000005821
【氏名又は名称】	松下電器産業株式会社
【代理人】	
【識別番号】	100097445
【弁理士】	
【氏名又は名称】	岩橋 文雄
【選任した代理人】	
【識別番号】	100103355
【弁理士】	
【氏名又は名称】	坂口 智康
【選任した代理人】	
【識別番号】	100109667
【弁理士】	
【氏名又は名称】	内藤 浩樹
【手数料の表示】	
【予納台帳番号】	011305
【納付金額】	16,000円
【提出物件の目録】	
【物件名】	特許請求の範囲 1
【物件名】	明細書 1
【物件名】	図面 1
【物件名】	要約書 1
【包括委任状番号】	9809938

【書類名】特許請求の範囲

【請求項 1】

構造化文書を管理するデータベース構築装置において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析手段と、
前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて要素名辞書に登録する要素名登録手段と、
前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて祖先パス名辞書に登録する祖先パス名登録手段と、
前記入力文書解析手段の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして要素出現情報格納手段に、
着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして祖先パス出現情報格納手段にそれぞれ登録する出現情報登録手段と、
を有することを特徴とするデータベース構築装置。

【請求項 2】

前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名 I D を割り当てて属性名辞書に登録する属性名登録手段と、
着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして格納した、属性出現情報格納手段とを有し、
前記出現情報登録手段が、前記入力文書解析手段の解析結果に基づいて、属性出現情報を作成して前記属性出現情報格納手段への登録も行うことを特徴とする請求項 1 に記載のデータベース構築装置。

【請求項 3】

要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と属性名 I D と分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納手段とを有し、
前記出現情報登録手段が、前記入力文書解析手段の解析結果に基づいて、テキスト出現情報を作成してテキスト出現情報格納手段にも登録することを特徴とする請求項 2 に記載のデータベース構築装置。

【請求項 4】

構造化文書を管理するデータベース検索装置において、
構造化文書に出現する各要素名に対してユニークな要素名 I D を登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして格納した要素出現情報格納手段と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして格納した、祖先パス出現情報格納手段と、
検索式を入力するための検索条件入力手段と、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換する検索条件解析手段と、
前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報及び、前記祖先パス出現情報格納手段からの祖先パス出現情報から検索結果

群を求める出現情報取得手段と、
を有することを特徴とするデータベース検索装置。

【請求項 5】

属性名 I D と対応する属性名の記録された属性名辞書と、
着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして格納した属性出現情報格納手段とを有し、
前記検索条件解析手段が、前記要素名辞書と前記祖先パス名辞書と前記属性名辞書とを参照して、前記検索条件入力手段から入力された検索式を内部条件式に変換し、前記出現情報取得手段が、前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報、前記祖先パス出現情報格納手段からの祖先パス出現情報及び、前記属性出現情報格納手段からの属性出現情報から検索結果群を求めることを特徴とする請求項 4 に記載のデータベース検索装置。

【請求項 6】

要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と属性名 I D と分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納手段とを有し、
前記出現情報取得手段が、前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報、前記祖先パス出現情報格納手段からの祖先パス出現情報、前記属性出現情報格納手段からの属性出現情報及び、前記テキスト出現情報格納手段からのテキスト出現情報から検索結果群を求めることを特徴とする請求項 5 に記載のデータベース検索装置。

【請求項 7】

前記出現情報取得手段は、前記要素出現情報格納手段における指定要素名 I D のエントリ数と、前記祖先パス出現情報格納手段における指定祖先パス名 I D のエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求めることを特徴とする請求項 4 乃至 6 のいずれかに記載のデータベース検索装置。

【請求項 8】

構造化文書を管理するデータベース構築方法において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行うステップと、
前記解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて要素名辞書に登録するステップと、
前記解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて祖先パス名辞書に登録するステップと、
前記解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして要素出現情報格納部に、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして祖先パス出現情報格納部にそれぞれ登録するステップと、を有することを特徴とするデータベース構築方法。

【請求項 9】

構造化文書を管理するデータベース検索方法において、
構造化文書に出現する各要素名に対してユニークな要素名 I D を登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして格納した要素出現情報格納部と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして

格納した、祖先パス出現情報格納部と、
検索式を入力するためのステップと、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式
に変換するステップと、
前記内部条件式に従って、前記要素出現情報格納部からの要素出現情報及び、前記祖先パ
ス出現情報格納部からの祖先パス出現情報から検索結果群を求めるステップと、
を有することを特徴とするデータベース検索方法。

【書類名】 明細書

【発明の名称】 データベース構築装置及びデータベース検索装置

【技術分野】

【0001】

本発明は、XMLなどの構造化文書を、管理するデータベース構築検索技術に関するもので、特に、大量の構造化文書群の中から指定した論理構造を持つ文書を効率良く検索するデータベース構築装置及びデータベース検索装置に関するものである。

【背景技術】

【0002】

従来の構造化文書を管理する装置としては、例えば特許文献1が知られている。

【0003】

以下、従来例の概要について図を参照しながら説明する。図24は従来の構造化文書管理装置の構成図である。登録対象の構造化文書は構造化文書入力手段2402から入力し、構造解析手段2407によって解析され、木構造を得る。構造情報作成手段2408によって、各要素のタグ名（要素名）には名称IDが割り振られて名称IDテーブル格納手段2418に格納される。また、各要素のパス名称（最上位階層から順にタグ名を連ねて記述した文字列）にはパス名称IDが割り振られて、パス名称インデックス格納手段2416に、各要素のパス階層（パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素か）を連ねて記述した文字列）にはパス階層IDが割り当てられて、パス階層インデックス格納手段2417に格納に、実体（テキスト）を持つ要素（要素実体）の場合は、各要素実体に対し、検索単位を一意に表す符合（検索単位識別子と呼ぶ）が割り当てられ、この検索単位識別子をキーとして、文書番号、パス名称ID、パス階層ID、名称IDの組が要素管理テーブル格納手段2415にそれぞれ格納される。図25は要素管理テーブルの例を示したものである。

【0004】

次に、文字列索引作成手段2409は、各要素実体の内容の文字列に対して、予め定めた文字数の文字連鎖を取り出す。この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（文字位置番号）を文字列索引格納手段2419に登録する。図26は文字列索引の例の一部を示した図である。図26の2601は「検索単位識別子が“1”の要素の文字列中に“構造”という文字連鎖が先頭から“1”文字目の位置から存在する」ということを表している。

【0005】

次に、このようにして格納されたデータを用いた検索の概要を説明する。図27は検索条件として「パス名称が“／論文／書誌／タイトル”である要素に“構造化”という文字列が含まれる文書」が与えられた場合の処理を図に示したものである。検索条件解析手段2410は、パス名称インデックス2416を参照し、検索条件のパス名称をパス名称ID“N2”に変換する。次に文字列索引検索手段2411は“構造化”から2文字連鎖“構造”と“造化”を取り出す。文字列索引を参照し、“構造”と“造化”が連続して出現し、かつ検索単位識別子が同一なものを求め、その検索単位識別子を抽出する。図27では検索単位識別子“1”と“8”が文字列索引検索結果群として返っている。次に、構造照合手段2412が検索条件の構造指定を満たす最終的な検索結果を求める。文字列索引検索結果群として得られた検索単位識別子をキーにして、要素管理テーブルを参照し、パス名称IDが“N2”に一致するものだけを最終的な検索結果とする。

【0006】

その他、タグ名を指定した検索条件であれば、要素管理テーブルの名称IDが指定タグ名の名称IDと一致するものだけを最終的な検索結果とし、パス名称とパス階層を共に指定した検索条件であれば、要素管理テーブルのパス名称IDが指定したパス名称のパス名称IDと一致し、かつパス階層IDが指定したパス階層のパス階層IDと一致するものだけを最終的な検索結果とする。

【特許文献1】 特開2002-202973号公報

【発明の開示】

【発明が解決しようとする課題】

【０００７】

しかしながら、前記従来の構成では、まず文字列索引を参照して指定された文字列の出現する検索単位識別子を求めた後、検索単位識別子が指定された構造条件を満たすかどうかを、要素管理テーブルを参照して判定する。そのため、文字列検索条件の指定は必須であり、構造条件だけを指定した検索を行うことができない（行うためには全ての検索単位識別子について構造条件を満たすかどうかを判定しなければならないため、効率が非常に悪い）という課題を有していた。また、文字列索引は要素実体の内容文字列に対してのみ作成されるため、属性値に対しては文字列検索を行うことができないという課題を有していた。

【０００８】

本発明は、前記従来の課題を解決するもので、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能なデータベース構築検索装置を提供することを目的とする。さらに、要素内のテキスト文字列だけでなく、属性値に対しても文字列検索が可能なデータベース構築検索装置を提供することを目的とする。

【課題を解決するための手段】

【０００９】

前記従来の課題を解決するために、本発明のデータベース構築検索装置は、要素の出現する文書番号、文字位置、文字数、祖先パス名ＩＤ、分岐順の情報を、要素名ＩＤをキーにして格納した要素出現情報格納手段と、要素の出現する文書番号、文字位置、文字数、要素名ＩＤ、分岐順の情報を、その要素の祖先パス名ＩＤをキーにして格納した祖先パス出現情報格納手段と、属性の出現する文書番号、文字位置、文字数、要素名ＩＤ、祖先パス名ＩＤ、分岐順の情報を、属性名ＩＤをキーにして格納した属性出現情報格納手段と、を備え、文字列検索条件を伴わない構造条件のみが指定された様々な検索式に対して、検索式の形と出現情報のエントリ数を考慮して、適切な出現情報を選択して使用することにより所望の文書を効率よく求めることができる。

【００１０】

また、要素実体のテキストから切り出した部分文字列、および要素のもつ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ＩＤ、要素名ＩＤ、属性名ＩＤ、分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納手段を備え、要素内のテキストに対する文字列検索だけでなく、属性値に対しても文字列検索が可能となる。

【発明の効果】

【００１１】

本発明のデータベース構築検索装置によれば、文字列検索条件と構造条件をともに指定した検索条件のみならず、構造だけを指定した様々な検索条件に対しても、所望の論理構造を持つ文書を効率よく検索することが可能となる。また、要素実体のテキスト文字列に対してだけでなく、属性値に対しても文字列検索を行うことが可能となる。

【発明を実施するための最良の形態】

【００１２】

以下本発明の実施の形態について、図面を参照しながら説明する。

【００１３】

（実施の形態１）

図１は、本発明の実施の形態１におけるデータベース構築検索装置の構成図である。

【００１４】

図１において、１０１はデータベースに登録する構造化文書群、１０２は入力された構造化文書群１０１の各文書についてユニークな文書番号を割り振るとともに論理構造の解析を行う入力文書解析手段、１０３は入力文書解析手段１０２の解析結果から、文書に出

現する要素名に対してユニークな識別子（以下、要素名 I D と呼ぶ）を割り当てて要素名辞書 1 0 7 に登録する要素名登録手段、1 0 4 は入力文書解析手段 1 0 2 の解析結果から、文書に出現する祖先パス名（着目要素の祖先要素の要素名を最上位階層から順にスラッシュで区切って並べた文字列で、着目要素自身の要素名は含まない）に対してユニークな識別子（以下、祖先パス名 I D と呼ぶ）を割り当てて祖先パス名辞書 1 0 8 に登録する祖先パス名登録手段、1 0 5 は入力文書解析手段 1 0 2 の解析結果から、文書に出現する属性名に対してユニークな識別子（以下、属性名 I D と呼ぶ）を割り当てて属性名辞書 1 0 9 に登録する属性名登録手段、1 0 6 は入力文書解析手段 1 0 2 の解析結果から、出現位置索引 1 1 0 の要素出現情報格納手段 1 1 1、祖先パス出現情報格納手段 1 1 2、属性出現情報格納手段 1 1 3、テキスト出現情報格納手段 1 1 4 に四種の出現情報を登録する出現情報登録手段、1 0 7 は要素名 I D とそれに対応する要素名が記録された要素名辞書、1 0 8 は祖先パス名 I D とそれに対応する祖先パス名が記録された祖先パス名辞書、1 0 9 は属性名 I D とそれに対応する属性名が記録された属性名辞書、1 1 0 は要素出現情報格納手段 1 1 1、祖先パス出現情報格納手段 1 1 2、属性出現情報格納手段 1 1 3、テキスト出現情報格納手段 1 1 4、の四種の出現情報が格納されている出現位置索引格納手段、1 1 1 は各要素の出現する文書番号、文字位置、文字数、祖先パス名 I D、分岐順の情報を、要素名 I D をキーにして格納した要素出現情報格納手段、1 1 2 は各要素の出現する文書番号、文字位置、文字数、要素名 I D、分岐順の情報を、その要素の祖先パス名 I D をキーにして格納した、祖先パス出現情報格納手段、1 1 3 は各属性の出現する文書番号、文字位置、文字数、要素名 I D、祖先パス名 I D、分岐順の情報を、属性名 I D をキーにして格納した属性出現情報格納手段、1 1 4 は要素内のテキストから切り出した部分文字列、および要素のもつ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名 I D、要素名 I D、属性名 I D、分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納手段、1 1 6 は検索式 1 1 5 を受け付ける検索条件入力手段、1 1 7 は、検索条件入力手段 1 1 6 に与えられた検索式を解析し、内部条件に変換して出現情報取得手段 1 1 8 に出力する検索条件解析手段、1 1 8 は検索条件解析手段 1 1 7 の出力した内部条件にしたがって、出現位置索引 1 1 0 に格納された四種の出現情報から適切な情報を選択して取得し、検索条件にマッチする結果データ集合を求める出現情報取得手段、1 1 9 は結果データ集合を適切な形式で検索結果 1 2 0 として出力する検索結果出力手段である。

【0 0 1 5】

上記のように構成されたデータベース構築検索装置の動作について説明する。

【0 0 1 6】

はじめに、文書登録（データベース構築）処理に関して具体例を挙げて説明する。図 2 は、文書の登録処理の流れを表す図である。

【0 0 1 7】

まず、ステップ 2 2 0 1 において、入力文書解析手段 1 0 2 は、構造化文書群 1 0 1 から構造化文書を 1 つ読み込んで、ユニークな文書番号を割り振る。

【0 0 1 8】

次に、ステップ 2 2 0 2 で、この文書の論理構造を解析する。図 2 は構造化文書の一例である（構造化文書群 1 0 1 には、このような文書が複数含まれる）。図 2 に示した構造化文書は、最上位階層に book 要素を持ち、book 要素は title 要素と 2 つの chapter 要素を含んでいる。title 要素は要素実体の文字列“文書検索”を含み、1 つ目の chapter 要素は別の title 要素と 2 つの section 要素および属性値が“歴史”である keyword 属性を持つ構造を持っている。図 2 に示す構造化文書を入力文書解析手段 1 0 2 によって解析した結果得られる木構造は、図 3 のようになる。図 3 において、四角い枠は要素 3 0 1 ～3 0 3 を表し、枠内に記された文字列は要素名 3 0 4 を示している。また、楕円の点線枠は属性 3 0 5 を表し、枠内に記された文字列は属性名 3 0 6 を示している。木構造の最上位階層の要素 3 0 1 から着目要素に至る経路の途中に存在する要素（祖先要素）の要素名をスラッシュで区切って順に並べたものはパス名と呼ばれる。パス名のうちの末尾部分（＝着目要素

自身の要素名)を除いた部分を「祖先パス名」と呼ぶことにする。図7は、図3の網掛けを施した要素302に関するパス名701, 祖先パス名702, 要素名703を示したものである。また、図3において、要素の右肩に記された“1/2/3”などの文字列は、パス名中の各要素について、同じ親要素を持つ同じ要素名の要素の中で何番目に出現したかの順を示す番号を並べたもので、これを「分岐順」307と呼ぶ。図3の網掛けを施した要素302とその左隣の要素303とは、パス名は同じであるが分岐順307, 308は異なっている。

【0019】

次に、入力文書解析手段102の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

【0020】

ステップ2203において、要素名登録手段103は、着目要素の要素名が要素名辞書107に登録済みかどうかを調べ、登録済みであれば対応する要素名IDを取得し、登録されていない場合は新たに要素名ID(>0)を割り当てて要素名辞書107に登録する。

【0021】

ステップ2204で、祖先パス名登録手段104は、着目要素の祖先パス名が祖先パス名辞書108に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名IDを取得し、登録されていない場合は新たに祖先パス名ID(>0)を割り当てて祖先パス名辞書108に登録する。

【0022】

もし、着目要素が属性を持っているならば、ステップ2205～ステップ2206において、属性名登録手段105は、着目要素の各属性の属性名が属性名辞書109に登録済みかどうかを調べ、登録済みであれば対応する属性名IDを取得し、登録されていない場合は新たに属性名ID(>0)を割り当てて属性名辞書109に登録する。図4、図5、図6はそれぞれ構造化文書(図2)の登録処理が終わった後の要素名辞書107、祖先パス名辞書108、属性名辞書109の内容の例を示している。

【0023】

ステップ2207において、出現情報登録手段106は、着目要素に関する要素出現情報を、要素名IDをキーとして要素出現情報格納手段111に登録する。要素出現情報は、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、祖先パス名ID、分岐順の5種類の値の組から構成される。なお、「文字位置」は、図8に示すように、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目に当たるかで表す。その一例を図9に示すもので、図3の網掛けを施した要素302に関する要素出現情報の内容で、要素名IDが4(=要素名がsection)である要素が文書番号1の文書の115文字目から始まる長さ40文字の要素実体を含んでいて、その祖先パス名IDが3(=祖先パス名が/book/chapter)で分岐順が1/2/3であることを表している。

【0024】

ステップ2208において、出現情報登録手段106は、着目要素に関する祖先パス出現情報(すなわち、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、要素名ID、分岐順の5種類の値の組)を、祖先パス名IDをキーとして祖先パス出現情報格納手段112に登録する。図10は、図3の網掛けを施した要素302に関する祖先パス出現情報の内容を示している。図9と図10を比較してわかるように、同一要素に関する要素出現情報と祖先パス出現情報は、キーとなる項目が要素名IDであるか祖先パス名IDであるかという点が異なるだけである。

【0025】

もし、着目要素が属性を持っているならば、ステップ2209～ステップ2210において、出現情報登録手段106は着目要素の各属性に関する属性出現情報を、属性名IDをキーとして属性出現情報格納手段113に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名ID、要素名ID、分岐順の6種類の値の組

から構成される。図 1 1 は、図 3 の網掛けを施した要素 3 0 2 の「update」属性 3 0 5 に関する属性出現情報の内容である。その内容は、属性名 I D が 2（＝属性名が update）の属性が文書番号 1 の文書の 1 1 5 文字目から始まる長さ 6 文字の属性値を持ち、属性の所属する要素の祖先パス名 I D が 3（＝祖先パス名が /book/section）、要素名 I D が 4（＝要素名が section）、分岐順が 1/2/3 であることを示している。なお、属性出現情報において、属性値の先頭文字位置は、図 1 1 に示すように、仮想的に着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置と同じであるとする。

【0 0 2 6】

ステップ 2 2 1 1 において、出現情報登録手段 1 0 6 は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納手段 1 1 4 に登録する。ただし、属性値ではないので、属性名 I D には常に 0 を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名 I D、要素名 I D、属性名 I D、分岐順の 6 種類の値の組から構成される。

【0 0 2 7】

もし、着目要素が属性を持っているならば、ステップ 2 2 1 2 ～ステップ 2 2 1 3 において、出現情報登録手段 1 0 6 は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納手段 1 1 4 に部分文字列をキーとして登録する。なお、属性出現情報と同様に、属性値は図 1 0 に示すような位置に仮想的に出現しているとして、文字位置を算出する。また、ステップ 2 2 1 2 ではステップ 2 2 1 1 の場合とは異なり、属性名 I D には着目している属性の属性名 I D（＞0）を格納する。図 1 2 は図 3 の網掛けを施した要素 3 0 2 のテキストおよび「update」属性 3 0 5 の属性値についてのテキスト出現情報の一部である。図 1 2 において、1 2 0 1 は、“極大”という部分文字列が文書番号 1 の文書の 1 1 8 文字目に現れ、祖先パス名 I D が 3（＝祖先パス名が /book/section）、要素名 I D が 4（要素名が chapter）、分岐順が 1/2/3 であるような要素の要素実体に含まれている（属性名 I D が 0 であることからわかる）ことを表している。また 1 2 0 2 は、“0 0”という部分文字列が文書番号 1 の文書の 1 1 8 文字目に現れ、祖先パス名 I D が 3（＝祖先パス名が /book/section）、要素名 I D が 4（＝要素名が chapter）、分岐順が 1/2/3 であるような要素に属する属性名 I D が 2（＝属性名が update）の属性の属性値に含まれていることを表している。

【0 0 2 8】

ステップ 2 2 1 4 で、この文書に出現する全ての要素について処理が終わったかどうかを調べ、もし未処理の要素が残っていればステップ 2 2 0 3 に戻って処理を繰り返す。

【0 0 2 9】

ステップ 2 2 1 5 で、全ての入力文書に対して処理が終わったかどうかを調べ、未処理の文書が残っていればステップ 2 2 0 1 に戻って処理を繰り返す。

【0 0 3 0】

以上のようにして、文書登録（データベース構築）処理が完了する。

【0 0 3 1】

続いて、登録済みの文書群に対する検索処理に関して説明する。

【0 0 3 2】

図 2 1 は、検索条件入力手段 1 1 6 に与えられる検索式 1 1 5 の例をいくつか示したもので、これらの式は W 3 C（World Wide Web Consortium）の勧告として公開されている X P a t h 言語（詳細な仕様は <http://www.w3.org/TR/xpath> に記載されている）で記述されている。

【0 0 3 3】

図 2 1 のそれぞれの X P a t h 式は、次のような意味を表している。検索式 2 1 0 1 は「最上位階層の book 要素の子の chapter 要素の子である title 要素」を、検索式 2 1 0 2 は「最上位階層の book 要素の子の chapter 要素のいずれかの子要素」を、検索式 2 1 0 3 は「いずれかの階層にある title 要素」を、検索式 2 1 0 4 は「最上位階層の book 要素の子

のchapter要素の子の2番目のsection要素」を、検索式2105は「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性」を、検索式2106は「最上位階層のbook要素の子のchapter要素の子のsection要素で、かつ要素実体内容に“極大単語”という文字列を含む要素」を、検索式2107は「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性で、かつその属性値に“2004”という文字列を含む」をそれぞれ表している。

【0034】

次に、それぞれの検索式に対して、どのような検索処理が行われるのかを順に説明する。

【0035】

(検索式2101の場合)

図23は、検索式2101の場合の検索処理の流れを示している。

【0036】

ステップ2301において、検索条件入力手段116に入力された検索式2101は、検索条件解析手段117で解析される。

【0037】

ステップ2302において、検索条件解析手段117は、検索式2101を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3かつ要素名ID=2」に変換し、出現情報取得手段118に出力する。

【0038】

次に、ステップ2303からステップ2305において、出現情報取得手段118は、出現位置索引格納手段110を参照し、要素出現情報格納手段111における要素名ID=2のエントリ数Nと祖先パス出現情報格納手段112における祖先パス名ID=3のエントリ数Mとを比較し、少ない方を選択する。図13は、要素出現情報格納手段111における要素名ID=2のエントリ1301、図14は祖先パス出現情報格納手段112における祖先パス名ID=3のエントリ1401の例で、この場合はN=8、M=12であるから図13の要素出現情報格納手段111を選ぶことになる。

【0039】

そして、ステップ2306において、要素出現情報格納手段111の要素名ID=2のエントリ1301から1つ取得し、ステップ2307で、このエントリの祖先パス名IDが3であるかどうかを調べ、もし祖先パス名IDが3であればステップ2308でこのエントリのデータを結果データ集合1302に追加する。結果データ集合の各データは例えば(文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順)のような形式である。

【0040】

ステップ2309でNエントリ全てについて処理したか調べ、まだ未処理のエントリがあればステップ2306に戻って処理を繰り返す。

【0041】

ステップ2305において、もし $M \leq N$ であれば、図14のように祖先パス出現情報格納手段112における祖先パス名ID=3の各エントリ1401を調べ、要素名IDが2であるものを求め(ステップ2310～ステップ2313)結果データ集合1402に追加する。

【0042】

ステップ2314で、求められた結果データ集合を検索結果出力手段119に出力する。最後に検索結果出力手段119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0043】

このように、検索式2101に対しては、要素出現情報格納手段111における指定した要素名IDのエントリから指定した祖先パス名IDを持つものを選ぶという処理と、祖先パス出現情報格納手段112における指定した祖先パス名IDのエントリから指定した要素名IDを持つものを選ぶという2種類の処理のどちらか、エントリ数の少ないほうを

選ぶことによって、検索対象構造化文書群の論理構造の特性に応じて処理量を抑えることができるため、所望の文書を効率よく検索することができる。

【0044】

（検索式2102の場合）

検索条件入力手段116に入力された検索式2102は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2102を解析し、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3」に変換し、出現情報取得手段118に出力する。出現情報取得手段118は、出現位置索引110を参照し、図15のように祖先パス出現情報格納手段112における祖先パス名ID=3の全てのエントリ1501を求め、例えば（文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順）のような形式で結果データ集合1502として検索結果出力手段119に出力する。検索結果出力手段119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0045】

このように、検索式2102に対しては、祖先パス出現情報格納手段112における指定した祖先パス名IDのエントリを取得するだけで良いため、所望の文書を効率よく検索することができる。

【0046】

（検索式2103の場合）

検索条件入力手段116に入力された検索式2103は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2103を解析し、要素名辞書107を参照して内部条件「要素名ID=2」に変換し、出現情報取得手段118に出力する。出現情報取得手段118は、出現位置索引格納手段110を参照し、図16のように要素出現情報格納手段111における要素名ID=2の全てのエントリ1601を求め、例えば（文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順）のような形式で結果データ集合1602を検索結果出力手段119に出力する。検索結果出力手段119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0047】

このように、検索式2103に対しては、要素出現情報格納手段111における指定した要素名IDのエントリを取得するだけで良いため、所望の文書を効率よく検索することができる。

【0048】

（検索式2104の場合）

検索条件入力手段116に入力された検索式2104は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2104を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3かつ要素名ID=4かつ分岐順=*/*/*2」に変換し、出現情報取得手段118に出力する。分岐順のアスタリスク「*」の部分はどんな数字でもマッチすることを表す。出現情報取得手段118は、出現位置索引110を参照し、要素出現情報格納手段111における要素名ID=4のエントリ数Nと祖先パス出現情報格納手段112における祖先パス名ID=3のエントリ数Mとを比較し、少ないほうを選択する。

【0049】

もし、 $M \leq N$ であれば、図17に示すように祖先パス出現情報格納手段112における祖先パス名ID=3の各エントリ1701を調べ、要素名IDが4であり、かつ分岐順が“*/*/*2”であるエントリのデータを結果データ集合1702として、例えば（文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順）のような形式で検索結果出力手段119に出力する。もし、 $M > N$ ならば要素出現情報格納手段111における要素名ID=4の各エントリを調べ、祖先パス名IDが3であり、かつ分岐順が“*/*/*2”であるエントリのデータを結果データ集合1702として検索結果出力手段119に出力する。

【0050】

最後に検索結果出力手段119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0051】

このように、検索式2104に対しては、要素出現情報格納手段111における指定した要素名IDのエントリから指定した祖先パス名IDと分岐順を持つものを選ぶという処理と、祖先パス出現情報格納手段112における指定した祖先パス名IDのエントリから指定した要素名IDと分岐順を持つものを選ぶという2種類の処理のどちらか、エントリ数の少ないほうを選ぶことによって、処理量を減らすことが可能となり、所望の文書を効率よく検索することができる。

【0052】

（検索式2105の場合）

検索条件入力手段116に入力された検索式2105は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2105を解析し、要素名辞書107、祖先パス名辞書108、属性名辞書109を参照して内部条件「祖先パス名ID=3かつ要素名ID=4かつ属性名ID=2」に変換し、出現情報取得手段118に出力する。出現情報取得手段118は、出現位置索引110を参照し、図18のように属性出現情報格納手段113における属性名ID=2の各エントリ1801を調べ、祖先パス名IDが3であり、要素名IDが4であればそのエントリのデータを例えば（文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順）のような形式で結果データ集合1802として検索結果出力手段119に出力する。最後に、検索結果出力手段119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0053】

このように、検索式2105に対しては、属性出現情報格納手段113における指定した属性名IDのエントリから指定した祖先パス名IDと要素名IDを持つものを選ぶことによって、所望の文書を検索することが可能となる。

【0054】

（検索式2106の場合）

検索条件入力手段116に入力された検索式2106は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2106を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3かつ要素名ID=4かつ要素内に文字列“極大単語”を含む」に変換し、出現情報取得手段118に出力する。出現情報取得手段118は、出現位置索引格納手段110を参照し、図19のようにテキスト出現情報格納手段114における“極大”のエントリ1901と“単語”のエントリ1902の間の接続演算を行う。その際、文書番号が同一であることと“単語”が“極大”の2文字後方に位置することだけでなく、祖先パス名IDが3、かつ要素名IDが4、かつ属性名IDが0、かつ分岐順が同一であるというチェックも行い条件を満たすものを、例えば（文書番号,祖先パス名ID,要素名ID,属性名ID,分岐順）のような形式で結果データ集合1903として検索結果出力手段119に出力する。検索結果出力手段119は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0055】

このように、検索式2106に対しては、テキスト出現情報格納手段114における部分文字列のエントリ同士の接続演算の際に、祖先パス名IDおよび要素名IDが指定した値であって、分岐順が同一であり、かつ属性名IDが0であるものを選ぶことによって、所望の文書を検索することが可能となる。

【0056】

（検索式2107の場合）

検索条件入力手段116に入力された検索式2107は、検索条件解析手段117で解析される。検索条件解析手段117は、検索式2107を解析し、要素名辞書107、祖

先パス名辞書 108、属性名辞書 109 を参照して内部条件「祖先パス名 ID=3 かつ要素名 ID=4 かつ属性名 ID=2 かつ属性値に文字列“2004”を含む」に変換し、出現情報取得手段 118 に出力する。出現情報取得手段 118 は、出現情報取得手段 118 は、出現位置索引 110 を参照し、図 20 のようにテキスト出現情報格納手段 114 における“20”のエントリ 2001 と“04”のエントリ 2002 の間の接続演算を行う。その際、文書番号が同一であることと“20”が“04”の 2 文字後方に位置することだけでなく、祖先パス名 ID が 3、かつ要素名 ID が 4、かつ属性名 ID が 2、かつ分岐順が同一であるというチェックも行い、条件を満たすものを、例えば（文書番号, 祖先パス名 ID, 要素名 ID, 属性名 ID, 分岐順）のような形式で結果データ集合 2003 として検索結果出力手段 119 に出力する。検索結果出力手段 119 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【0057】

このように、検索式 2107 に対しては、テキスト出現情報格納手段 114 における部分文字列のエントリ同士の接続演算の際に、祖先パス名 ID および要素名 ID が指定した値であって、分岐順が同一であり、かつ属性名 ID が指定した値(>0)であるものを選ぶことによって、所望の文書を検索することが可能となる。

【0058】

以上説明したように、要素の出現情報を、要素名 ID をキーにして格納した要素出現情報格納手段と、要素の出現情報をその要素の祖先パス名 ID をキーにして格納した祖先パス名出現情報格納手段と、属性の出現情報を、属性名 ID をキーにして格納した属性出現情報格納手段とを設けることにより、構造条件だけを指定した検索式に対しても効率よく所望の文書を検索することができる。また、要素実体のテキスト文字列および要素の持つ属性の属性値から切り出された部分文字列の出現情報を格納したテキスト出現情報格納手段を設けることにより、要素実体のテキストに対してだけでなく属性値に対しても文字列検索を行うことができる。

【0059】

なお、データベース構築処理において、要素実体や属性値から固定長の 2 文字連鎖で部分文字列の切り出しを行うと説明したが、他の切り出し方法、例えば特開平 8-249354 号公報「文書検索装置および単語索引作成方法および文書検索方法」に記載の方法等でも構わない。

【0060】

また、データベース検索処理において、検索条件式を X P a t h 式で与えるとして説明したが、同様の意味を持つ他のクエリ言語であっても本発明を適用することは可能である。

【産業上の利用可能性】

【0061】

本発明のデータベース構築検索装置は、技術文献、特許文献、新聞、雑誌等、あらゆる電子化された構造化文書に対して広範に利用することが可能である。

【図面の簡単な説明】

【0062】

【図 1】 本発明の実施の形態 1 の構成を示すブロック図

【図 2】 本発明の実施の形態 1 における、登録検索対象となる構造化文書の一例を示す図

【図 3】 本発明の実施の形態 1 における、構造化文書の論理構造を解析した結果である木構造の一例を示す図

【図 4】 発明の実施の形態 1 における、要素名辞書の内容の一例を示す図

【図 5】 本発明の実施の形態 1 における、祖先パス名辞書の内容の一例を示す図

【図 6】 本発明の実施の形態 1 における、属性名辞書の内容の一例を示す図

【図 7】 本発明の実施の形態 1 における、祖先パス名の説明に用いる図

【図 8】 本発明の実施の形態 1 における、文字位置の説明に用いる図

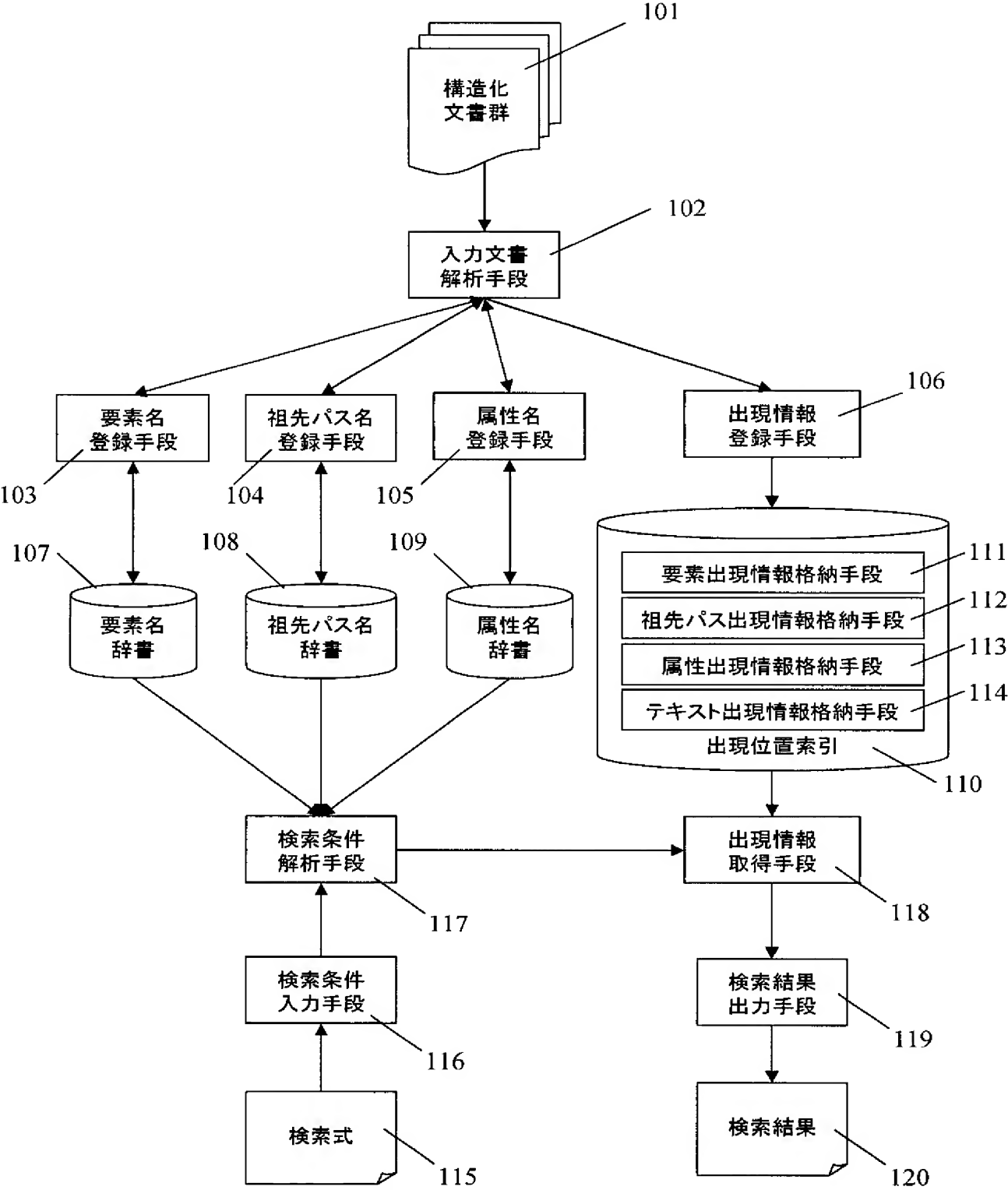
- 【図 9】 本発明の実施の形態 1 における、要素出現情報の説明に用いる図
- 【図 10】 本発明の実施の形態 1 における、祖先パス出現情報の説明に用いる図
- 【図 11】 本発明の実施の形態 1 における、属性出現情報の説明に用いる図
- 【図 12】 本発明の実施の形態 1 における、テキスト出現情報の説明に用いる図
- 【図 13】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 14】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 15】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 16】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 17】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 18】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 19】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 20】 本発明の実施の形態 1 における、検索処理の説明に用いる図
- 【図 21】 本発明の実施の形態 1 における、検索式の例を示す図
- 【図 22】 本発明の実施の形態 1 における、文書登録処理の手順を示す流れ図
- 【図 23】 本発明の実施の形態 1 における、検索処理の手順を示す流れ図
- 【図 24】 従来の技術における構造化文書管理装置の構成図
- 【図 25】 従来の技術における要素管理テーブルの例を示す図
- 【図 26】 従来の技術における文字列索引の例の一部を示す図
- 【図 27】 従来の技術における検索処理の図

【符号の説明】

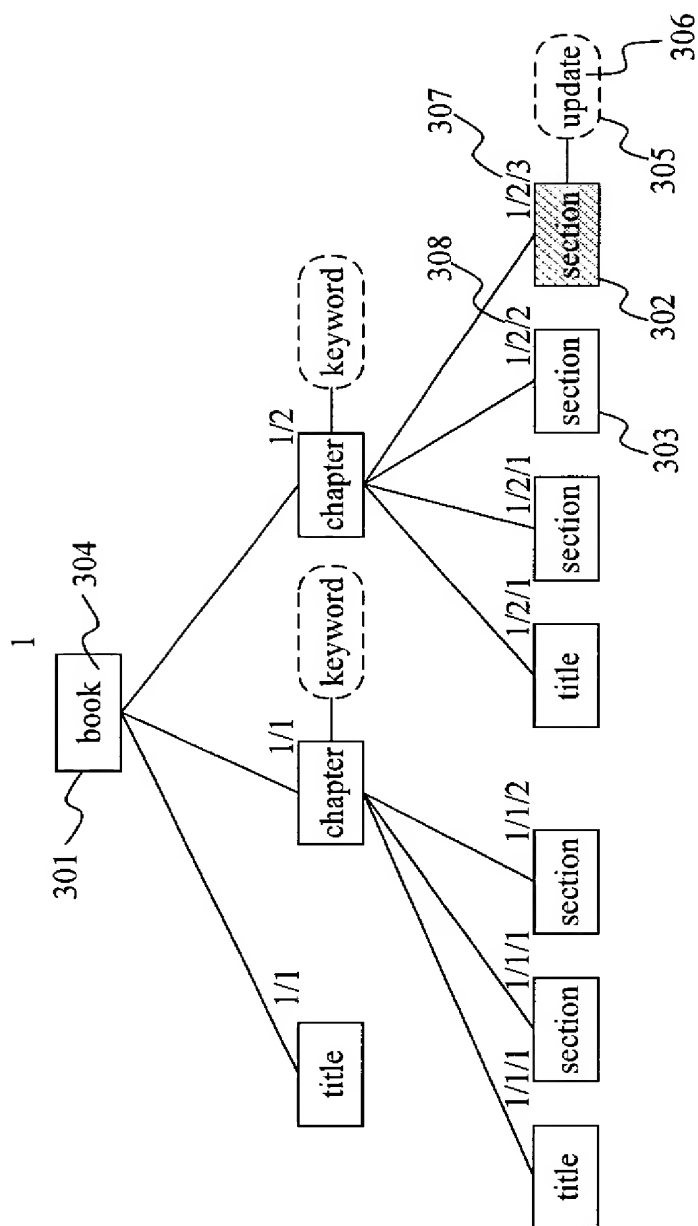
【0063】

- 101 構造化文書群
- 102 入力文書解析手段
- 103 要素名登録手段
- 104 祖先パス名登録手段
- 105 属性名登録手段
- 106 出現情報登録手段
- 107 要素名辞書
- 108 祖先パス名辞書
- 109 属性名辞書
- 110 出現位置索引
- 111 要素出現情報格納手段
- 112 祖先パス出現情報格納手段
- 113 属性出現情報格納手段
- 114 テキスト出現情報格納手段
- 115 検索式
- 116 検索条件入力手段
- 117 検索条件解析手段
- 118 出現情報取得手段
- 119 検索結果出力手段
- 120 検索結果

【書類名】 図面
【図 1】



```
<book>
  <title>文書検索</title>
  <chapter keyword="歴史">
    <title>文書検索の歴史</title>
    <section>キーワード検索は、・・・</section>
    <section>その後、全文検索が・・・</section>
  </chapter>
  <chapter keyword="索引">
    <title>索引方式</title>
    <section>最長一致切り出しによる・・・</section>
    <section>n-gram索引方式は・・・</section>
    <section update="200406">新たに極大単語索引方式が・・・</section>
  </chapter>
</book>
```



【 図 4 】

要素名ID	要素名
1	book
2	title
3	chapter
4	section

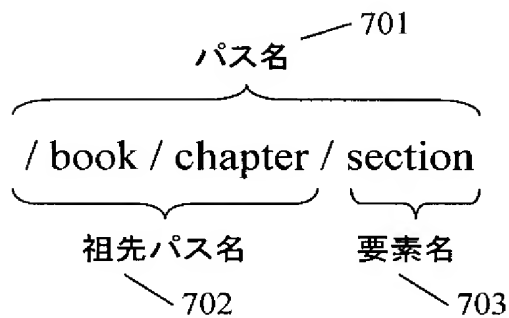
【図 5】

祖先パス名ID	祖先パス名
1	/
2	/book
3	/book/chapter

【図 6】

属性名ID	属性名
1	keyword
2	update

【図 7】

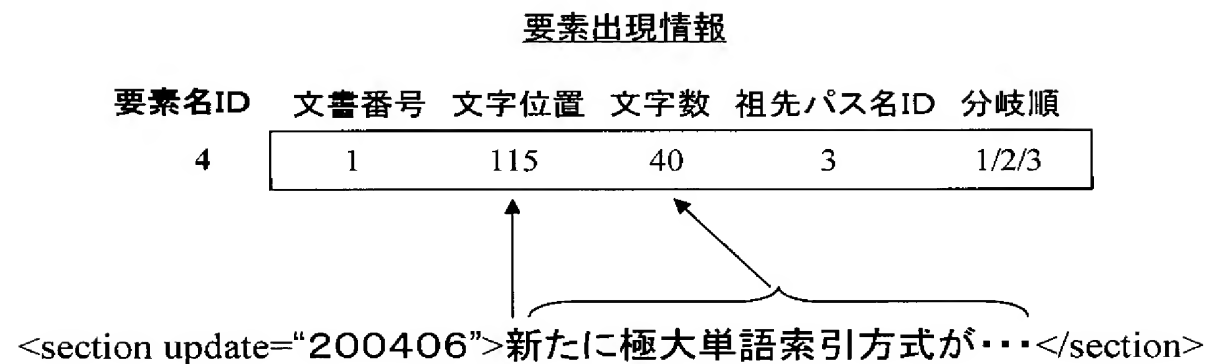


【図 8】

テキスト → 文 書 検 索 文 書 検 索 の 歴 史
 文字位置 → 0 1 2 3 4 5 6 7 8 9 10

キ ー ワ ー ド 検 索 は 、 . . .
 11 12 13 14 15 16 17 18 19

【図 9】



【図 1 0】

祖先パス出現情報

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
3	1	115	40	4	1/2/3

【図 1 1】

属性出現情報

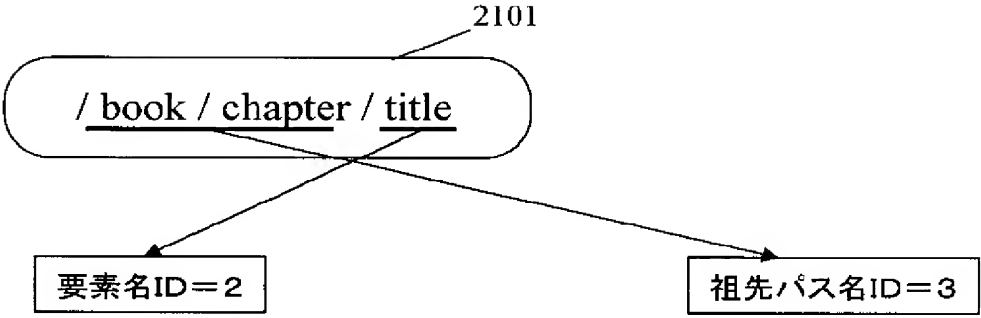
属性名ID	文書番号	文字位置	文字数	祖先パス名ID	要素名ID	分岐順
2	1	115	6	3	4	1/2/3



テキスト出現情報

部分文字列	文書番号	文字位置	祖先パス名ID	要素名ID	属性名ID	分岐順
1201 — “新た”	1	115	3	4	0	1/2/3
“たに”	1	116	3	4	0	1/2/3
“に極”	1	117	3	4	0	1/2/3
“極大”	1	118	3	4	0	1/2/3
“大単”	1	119	3	4	0	1/2/3
“単語”	1	120	3	4	0	1/2/3
<div>属性値でない場合は0 属性値の場合は属性名ID(≠0)</div>						
1202 — “2 0”	1	115	3	4	2	1/2/3
“0 0”	1	116	3	4	2	1/2/3
“0 4”	1	117	3	4	2	1/2/3
“4 0”	1	118	3	4	2	1/2/3
“0 6”	1	119	3	4	2	1/2/3

【図 1 3】



要素出現情報

要素名ID 文書番号 文字位置 文字数 祖先パス名ID 分岐順

2	.				
	.				
	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
3	7	0	5	2	1/1
	9	7	4	3	1/1/1
	.				
.					
.					

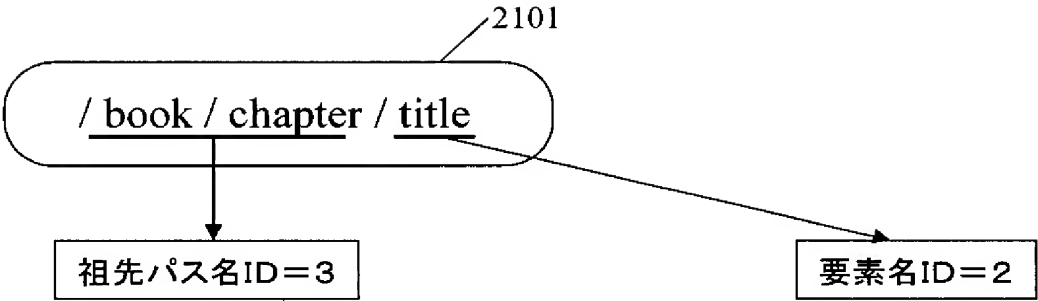
1301

8エントリ

結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={ (1, 3, 2, 0, 1/1/1),
 (1, 3, 2, 0, 1/2/1),
 (4, 3, 2, 0, 1/1/1),
 (9, 3, 2, 0, 1/1/1) }

1302



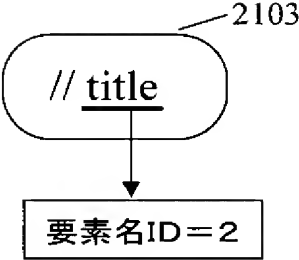
祖先パス出現情報

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4					

結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={ (1, 3, 2, 0, 1/1/1),
(1, 3, 2, 0, 1/2/1),
(4, 3, 2, 0, 1/1/1),
(9, 3, 2, 0, 1/1/1) }

【図 1 6】



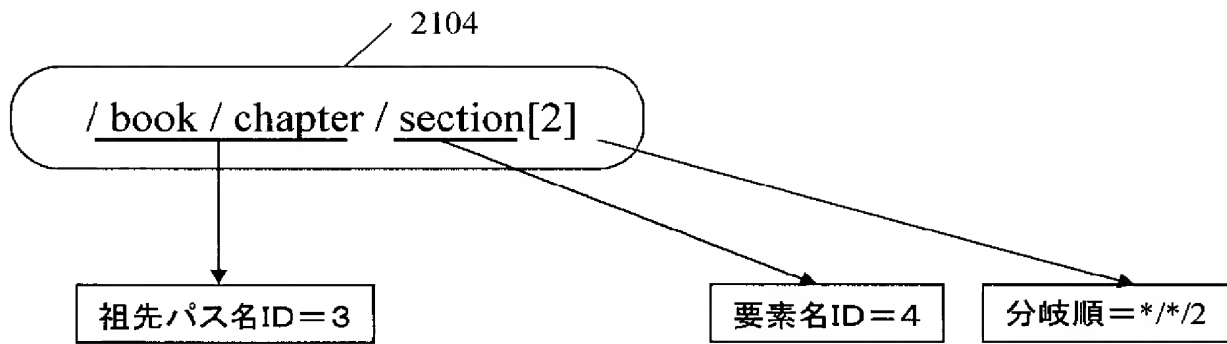
要素出現情報

要素名ID 文書番号 文字位置 文字数 祖先パス名ID 分岐順

2					
	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
3	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1

結果データ集合

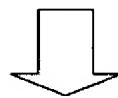
(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={ (1, 2, 2, 0, 1/1),
 (1, 3, 2, 0, 1/1/1),
 ...
 (7, 2, 2, 0, 1/1),
 (9, 3, 2, 0, 1/1/1) }



祖先パス出現情報

祖先パス名ID 文書番号 文字位置 文字数 要素名ID 分岐順

3						
	1	4	7	2	1/1/1	1701
	1	11	28	4	1/1/1	
	1	39	20	4	1/1/2	
	1	59	4	2	1/2/1	
	1	63	30	4	1/2/1	
	1	93	22	4	1/2/2	
	1	115	40	4	1/2/3	
	3	5	25	4	1/1/1	
	4	24	6	2	1/1/1	
	4	60	15	4	1/1/1	
	6	64	6	4	1/1/1	
	9	7	4	2	1/1/1	
4						

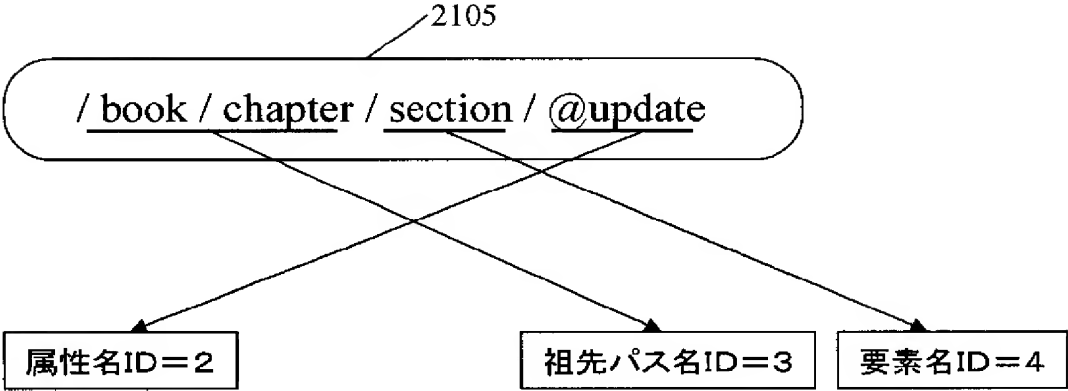


結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
 = { (1, 3, 4, 0, 1/1/2),
 (1, 3, 4, 0, 1/2/2) }

1702

【図 18】



属性出現情報

属性名ID 文書番号 文字位置 文字数 祖先パス名ID 要素名ID 分岐順

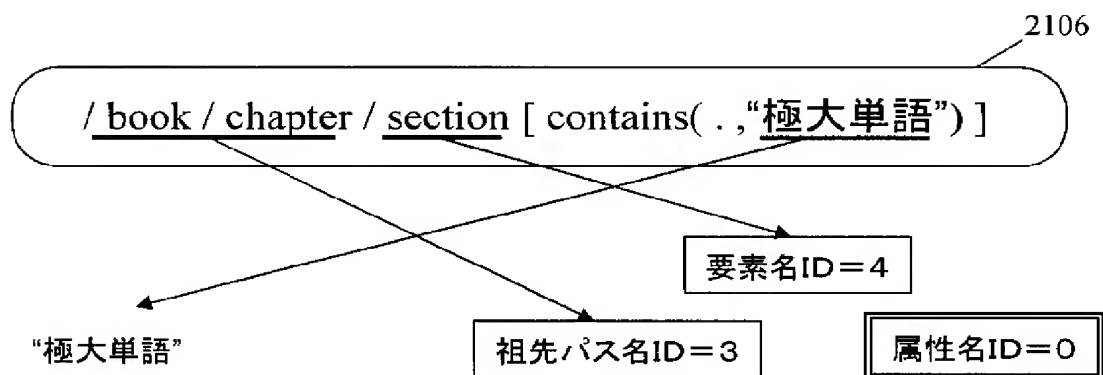
2	⋮					
	1	115	6	3	4	1/2/3
	2	8	4	2	2	1/1
	5	60	6	3	4	1/1/2
	8	32	8	3	2	1/2/1
3	⋮					
	⋮					

1801

結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={ (1, 3, 4, 2, 1/2/3),
 (5, 3, 4, 2, 1/1/2) }

1802



テキスト出現情報

部分文字列 文書番号 文字位置 祖先パス名ID 要素名ID 属性名ID 分岐順

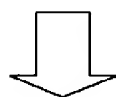
“極大”	1	118	3	4	0	1/2/3
	2	86	3	4	0	1/1/1
	3	24	2	2	0	1/1
	4	62	3	4	0	1/1/1
	8	77	3	4	2	1/1/1
			.			
			.			
			.			

1901

文書番号が同じで文字位置が接続、分岐順も同じであること

“単語”	1	120	3	4	0	1/2/3
	3	26	2	2	0	1/1
	4	64	3	4	0	1/1/1
	8	79	3	4	1	1/1/1
			.			
			.			
			.			

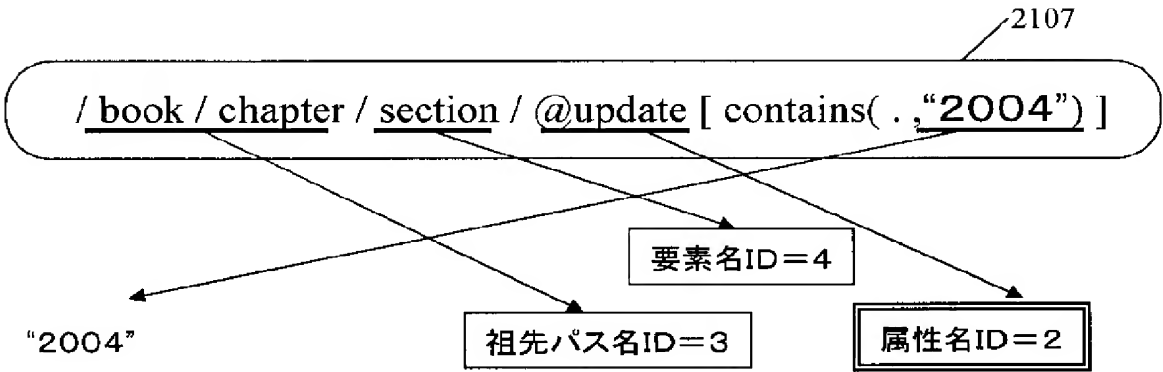
1902



結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
 = { (1, 3, 4, 0, 1/2/3),
 (4, 3, 4, 0, 1/1/1) }

1903



テキスト出現情報

部分文字列 文書番号 文字位置 祖先パス名ID 要素名ID 属性名ID 分岐順

"20"

1	115	3	4	2	1/2/3
2	15	3	4	0	1/1/1
3	24	2	2	0	1/1
5	21	3	4	2	1/1/1
7	54	3	4	1	1/1/1
⋮					

2001

文書番号が同じで文字位置が接続、分岐順も同じであること

"04"

1	117	3	4	2	1/2/3
3	26	2	2	0	1/1
5	23	3	4	2	1/1/1
7	56	3	4	1	1/1/1
⋮					

2002



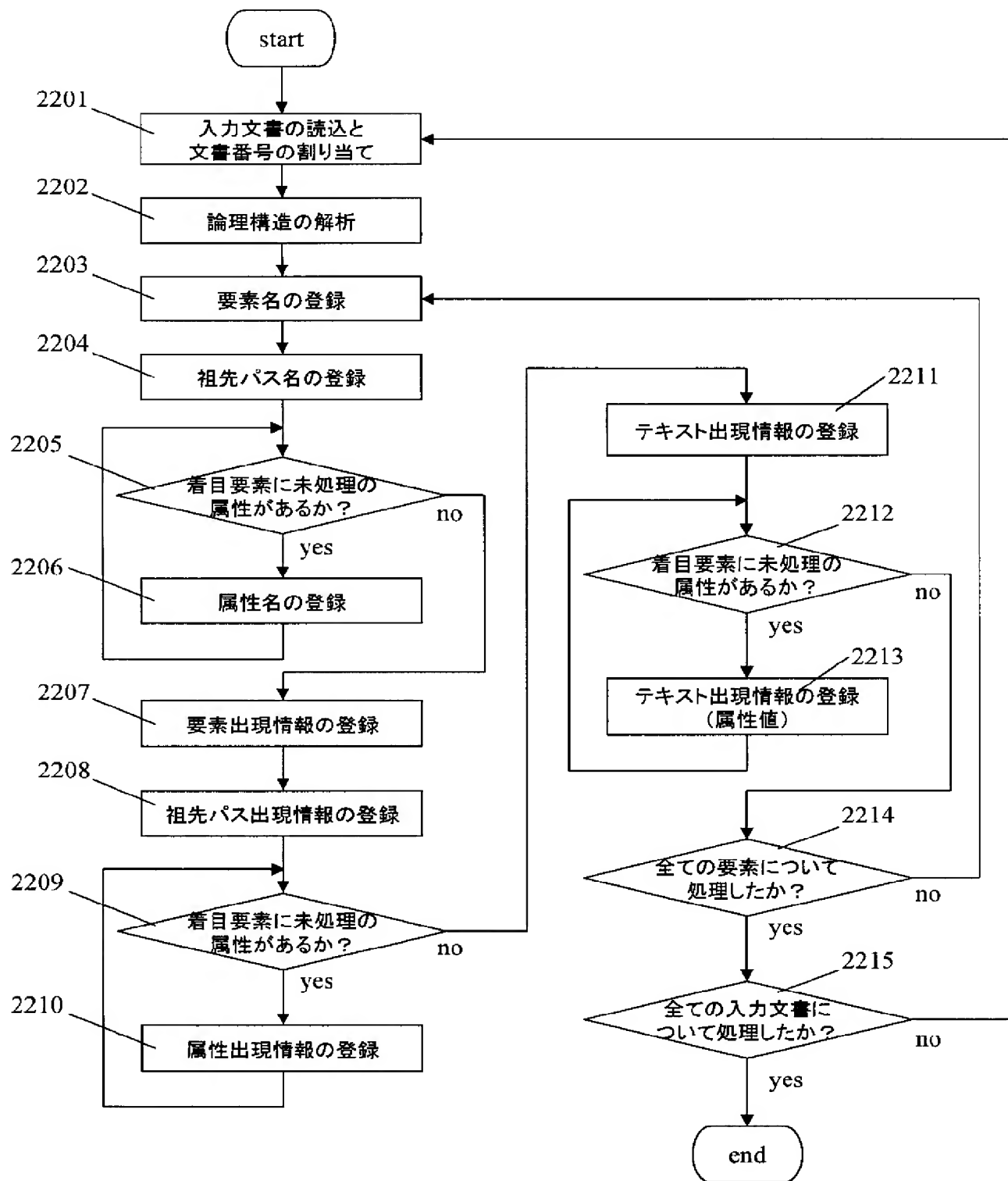
結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={ (1, 3, 4, 2, 1/2/3),
 (5, 3, 4, 2, 1/1/1) }

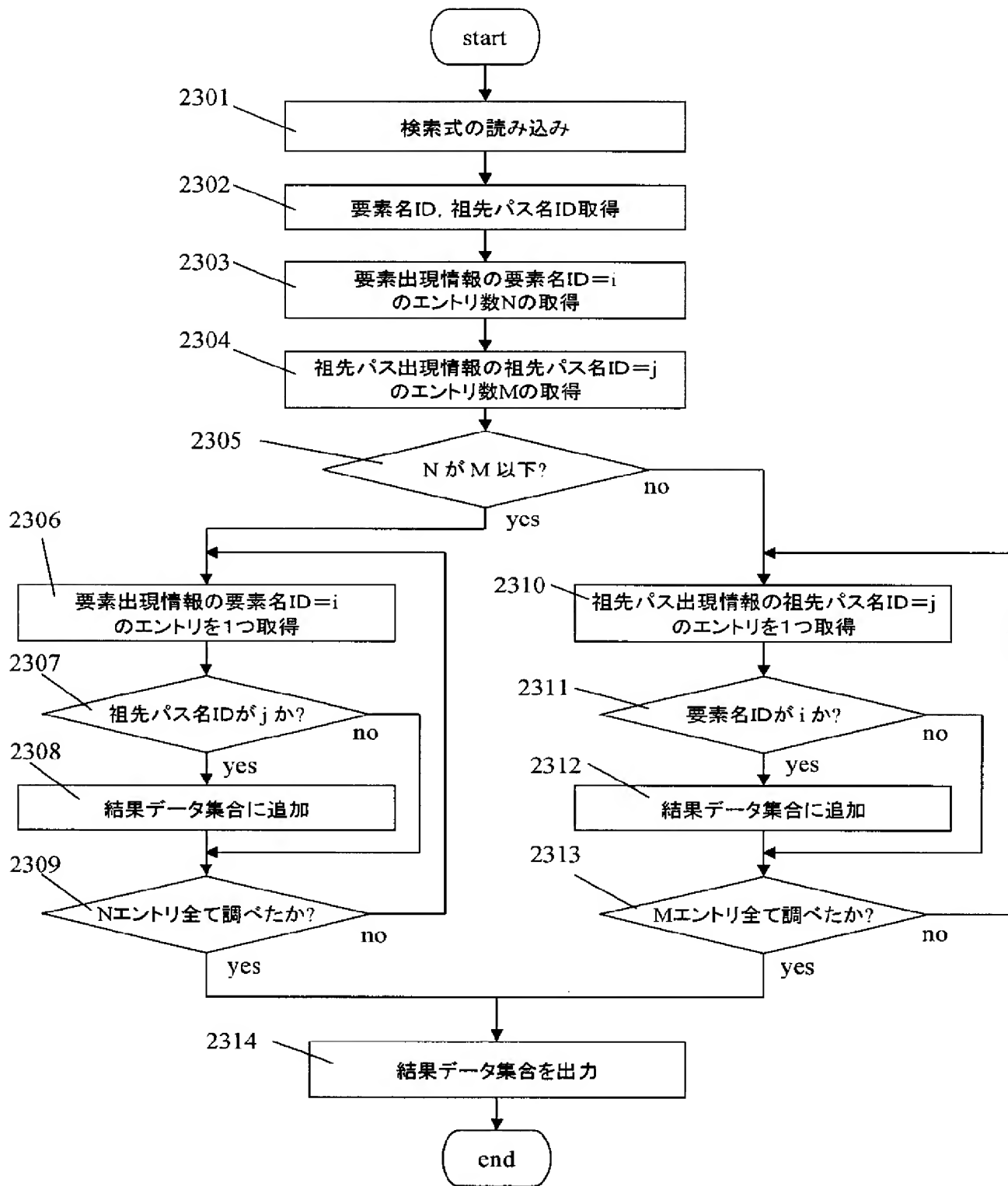
2003

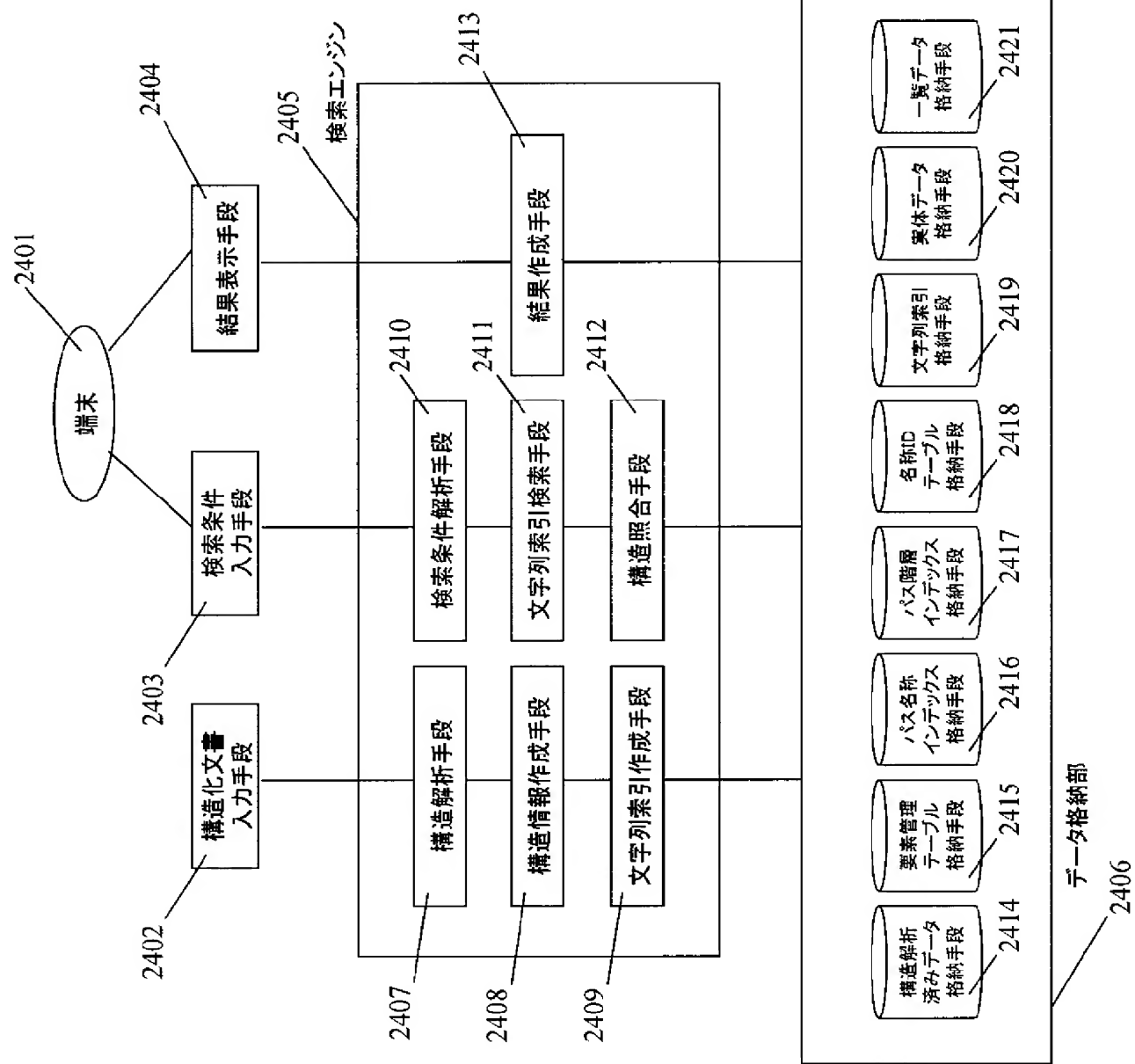
```
2101 /book / chapter / title
2102 /book / chapter / *
2103 // title
2104 /book / chapter / section[2]
2105 /book / chapter / section / @update
2106 /book / chapter / section [ contains( . , “極大単語”) ]
2107 /book / chapter / section / @update [ contains( . , “2004”) ]
```

【図 2 2】



【図 2 3】





【図 2 5】

検索単位
識別子 文書番号 パス名称ID パス階層ID 名称ID

1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L6	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
.
.				
.				

要素管理テーブル

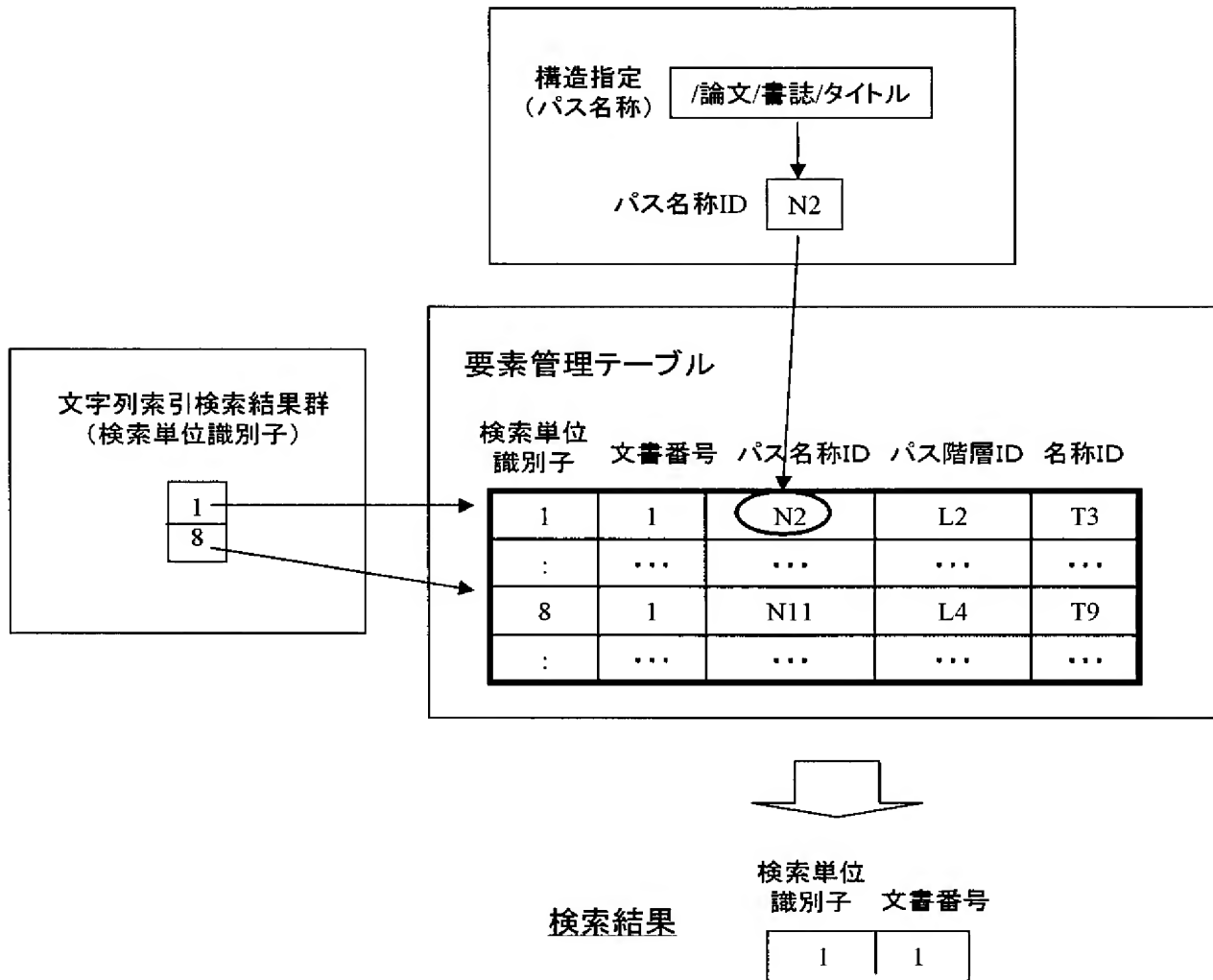
【図 2 6】

<タイトル>構造化文書管理</タイトル>

“構造”	1	1	2601
“造化”	1	2	
“化文”	1	3	
“文書”	1	4	
“書管”	1	5	
“管理”	1	6	

検索単位
識別子

文字位置番号



【書類名】 要約書

【要約】

【課題】 様々な検索条件での構造化文書に対する検索を効率よく行う。構造条件のみでの検索や、属性値に対する文字列検索ができるようにすることを目的とする。

【解決手段】 要素の出現情報を、要素名 I D をキーにして格納した要素出現情報格納手段と、要素の出現情報を、その要素の祖先パス名 I D をキーにして格納した祖先パス出現情報格納手段と、属性の出現情報を、属性名 I D をキーにして格納した属性出現情報格納手段と、要素実体のテキスト文字列、および要素のもつ属性の属性値に関する出現情報を、部分文字列をキーにして格納したテキスト出現情報格納手段とを備える。

【選択図】 図 1

出願人履歴

0 0 0 0 0 5 8 2 1

19900828

新規登録

大阪府門真市大字門真 1 0 0 6 番地

松下電器産業株式会社